Summarizing geo-annotations from social media: a promising approach for smart cities and communities

Rosa Meo, Ruggero G. Pensa, Mattia Bertorello and Gianpiero Di Blasi

Dipartimento di Informatica, University of Torino

UNIVERSITÀ DEGLI STUDI DI TORINO

ALMA UNIVERSITAS TAURINENSIS

SMATF2

Motivations

Background: Social media services generate huge amounts of (open) up-to-date geo-referenced data (e.g., Facebook, Twitter, Foursquare)

Aim:

provide a **first glance summary** of the geographical extent of a region of interest (in this work we focus on Milan and Trentino regions)

Geosummly

1. An application for **collecting** and **clustering** venues coming from crowdsourcing applications
2. Provides a **first glance summary** of the geographical extent of a region of interest (here we focus on Milan and Trentino regions)

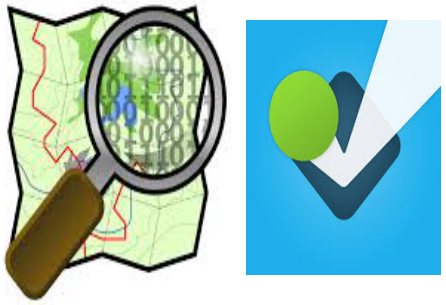
SMATF2

Domains of application of Geosummly

Geosummly has the typical applications of **smart city**:

- Representation and exploratory analysis of geographical features
- Integration of annotations from multiple sources
- City security
- Mobility
- Tourism
- Social policy making
- Environment monitoring
-

Geosummly Overview



Geosummly automatically adds a layer over a typical geographic maps service (e.g., geoServer, OpenStreetMap), with **fingerprints**

A fingerprint is a geographic summary of the venues extracted from the social media geographic annotation platforms (Foursquare)

A fingerprint is represented by a colored polygon whose shape is automaltically computed by **convex hull** on the venues in the area.

Each fingerprint is **labeled with the most prominent categories.**

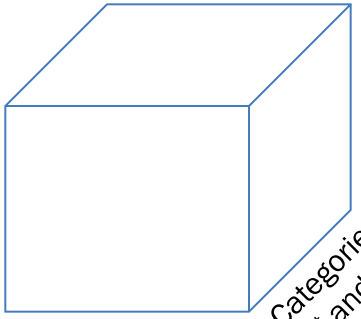
Example:
Food, Travel & Transport.

Demo: http://geosummly.eurecom.fr

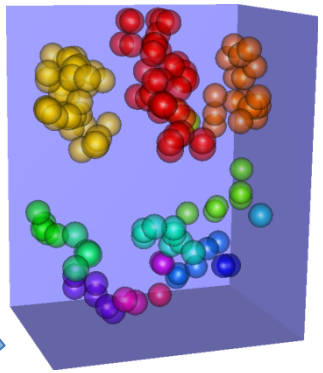
SMATF2

Pipeline

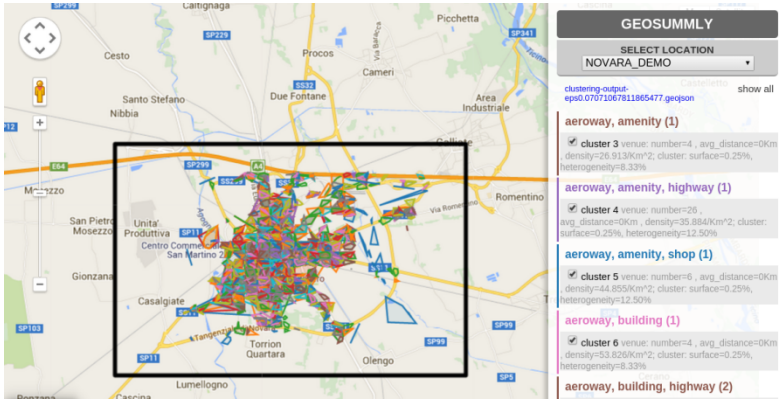
Sampling and fusion



Density Based Clustering Algorithm

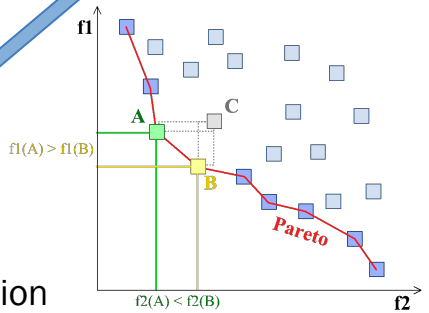


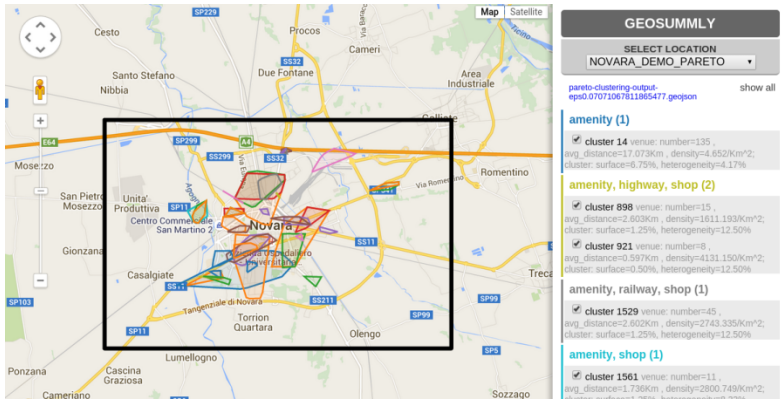
Geographic summaries (GeoJSON)



Visualization and publication (Open Annotation Model)

Pareto Optimization (multistrategy)





SMATF2

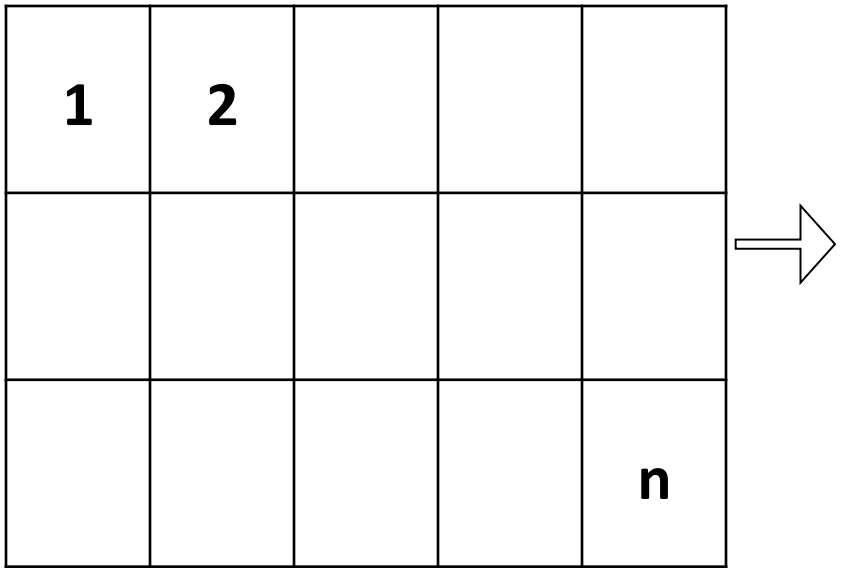
Idea

Description

- For each cell of a given grid on a geographical area, we collect the Foursquare venues,
  - compute the venue distribution for each category (e.g. Food, Travel & Transport, ...).
  - We consider both the 1st and the 2nd level of the categories of the OpenStreetMap and Foursquare taxonomy
- We build a N dimensional model
  - the first two dimensions are latitude, longitude
  - the others represent the categories from the social media taxonomy
- We apply a density based cluster analysis.
  - The clustering algorithm aggregates cells that are spatially close in which the density of each given category is uniform.
- For each cluster, we compute a geographic shape
- We publish on LinkedOpenData (LOD) and the Open Annotation Ontology
- The clusters are filtered by a Pareto optimization function

Grid Sampling on Foursquare

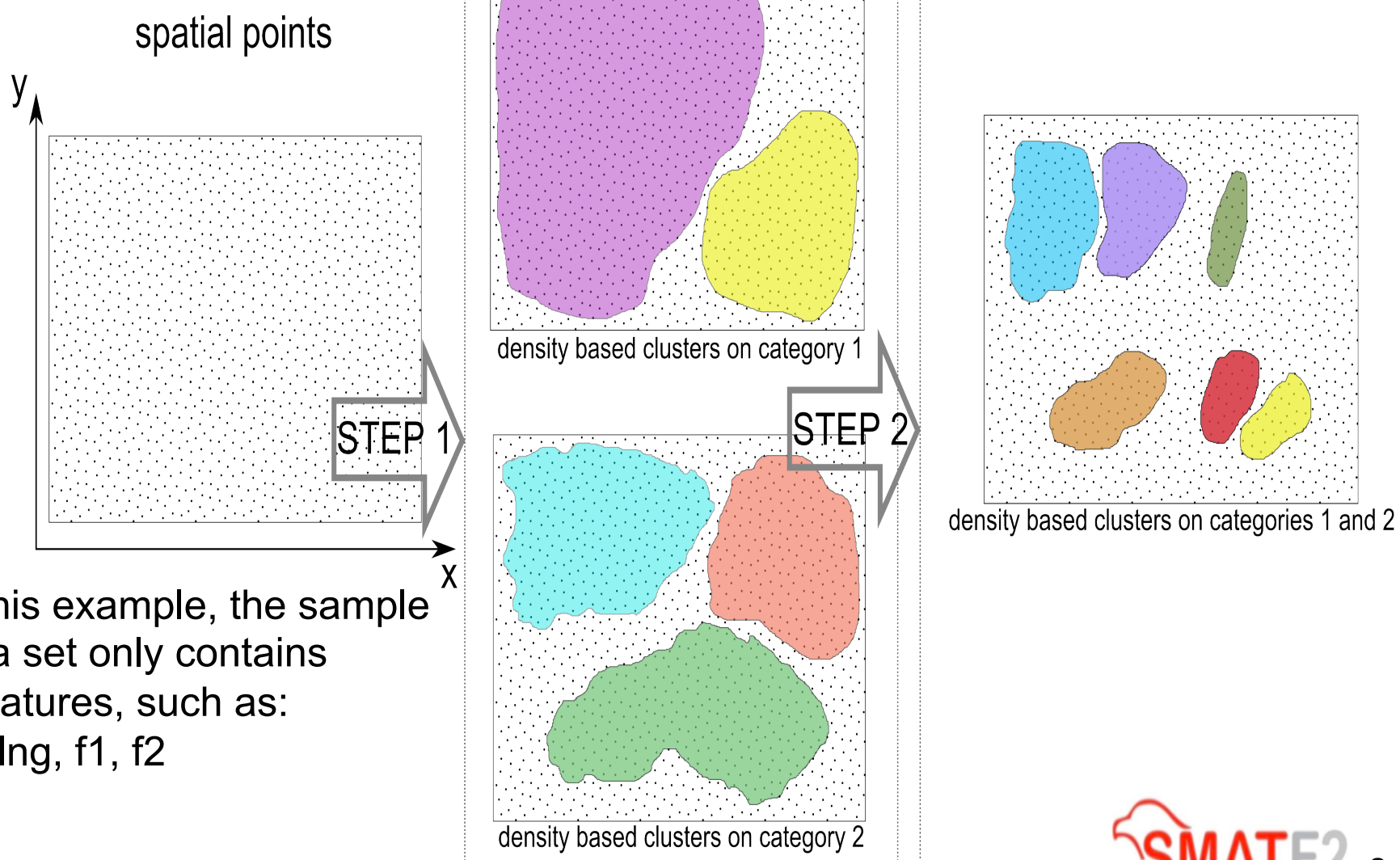
Grid



			Arts & Entertainment	College & University	
1	lat*	lng*	f(1,1)	f(1,2)	
2					
3					
...					
n					f(n,n)

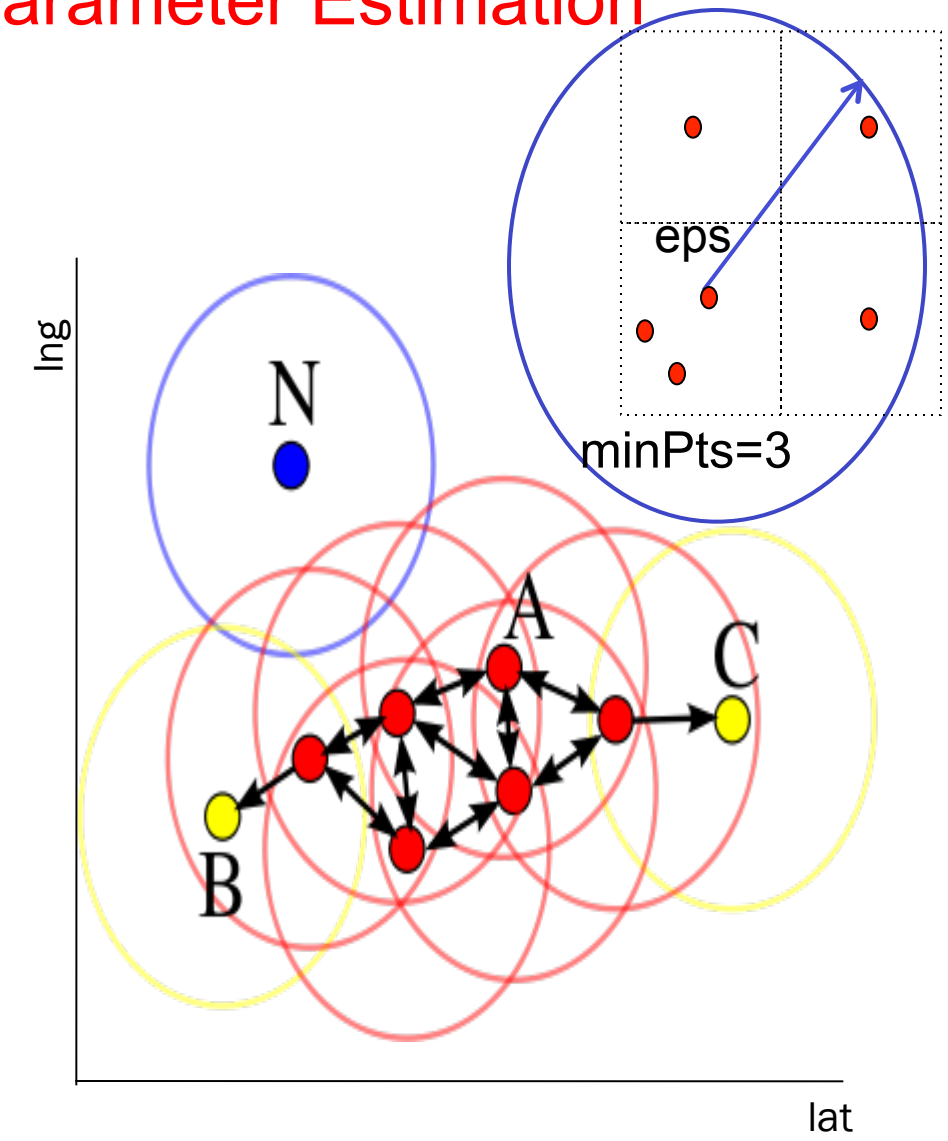
lat*,long* = latitude and longitude of the centroid of the cell(i).
Reduces the observation noise of the single venues, and to reduce the data set sparsity.

Clustering with GeoSubClu in action



In this example, the sample data set only contains 4 features, such as: lat, lng, f1, f2

Parameter Estimation



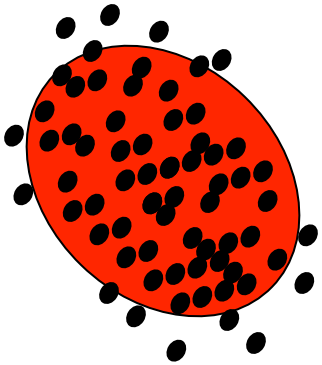
Definitions:

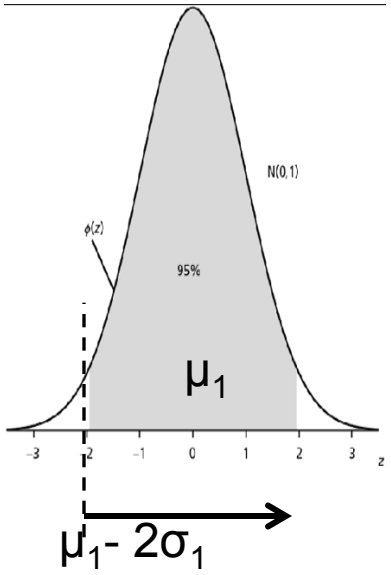
- **eps :** reachable distance. We use the Euclidean distance (points linked with arrows)
- **minPts :** min number of points to have a cluster (given the example, it can be 1...8)

Automatic parameter estimation in a multidimensional space:

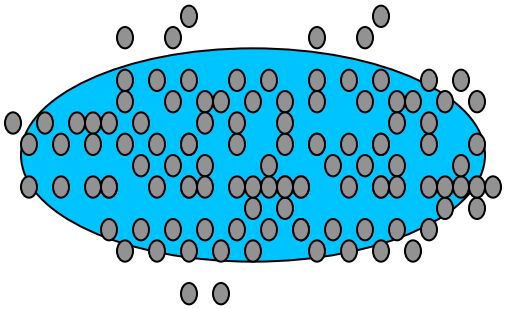
- **eps** : applying the Euclidean distance in a dimensional space which is progressively larger (the number of considered features)
- **minPts:**
  We distinguish between single features and combined features
- **single:** considering each feature separately, minPts= mean frequency value observed in the entire grid, reduced by an amount that allows to eliminate the lower frequencies (outliers)
- **combined:** we assume features are independent; so are their frequencies (random variables);
  we compute their product

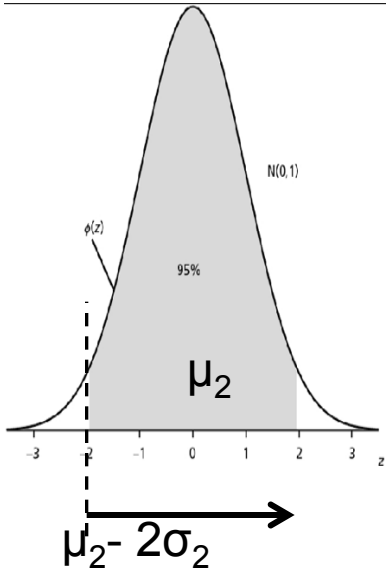
Automatic parameter estimation: combined features





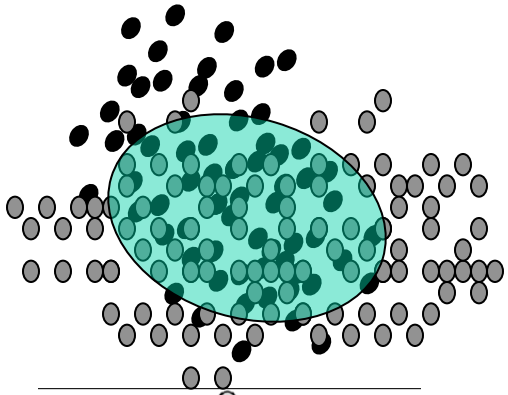
$\mu_1 - 2\sigma_1$

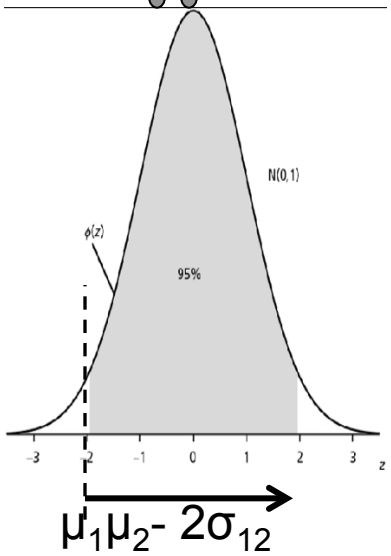




$\mu_2 - 2\sigma_2$

Independence assumption





$\mu_1\mu_2 - 2\sigma_{12}$

$$\sigma_{12} = \sqrt{\frac{\sum_{cells\_i} (f_{i1} \cdot f_{i2} - \mu_1 \cdot \mu_2)^2}{n\_cells}}$$

Cluster Validation: Main Outcomes

First glance summaries of two areas (Milan and Trentino) from the venue categories of Foursquare users

Validation:

.clustering-based (SSE)

Compared the Sum of Squared Errors of clusters in real data and in 500 random data

As a result, we have **only 3% chance that clusters occur by chance**

.geographic summary-based (Jaccard)

We applied the **hold out** (ratio 50%), resulting in two data sets: A, B
We applied Geosummly to A and B and measured the overlap of the categories in the clusters using the Jaccard distance.
We observed an **average overlap of the fingerprints of 82%** (70% is the minimal overlap, statistically assessed).

Validation by a panel of crowdsourcing applications

Clusters correspond with a high precision (~70-80%) to perceived prominent categories

Optimization: evaluation functions

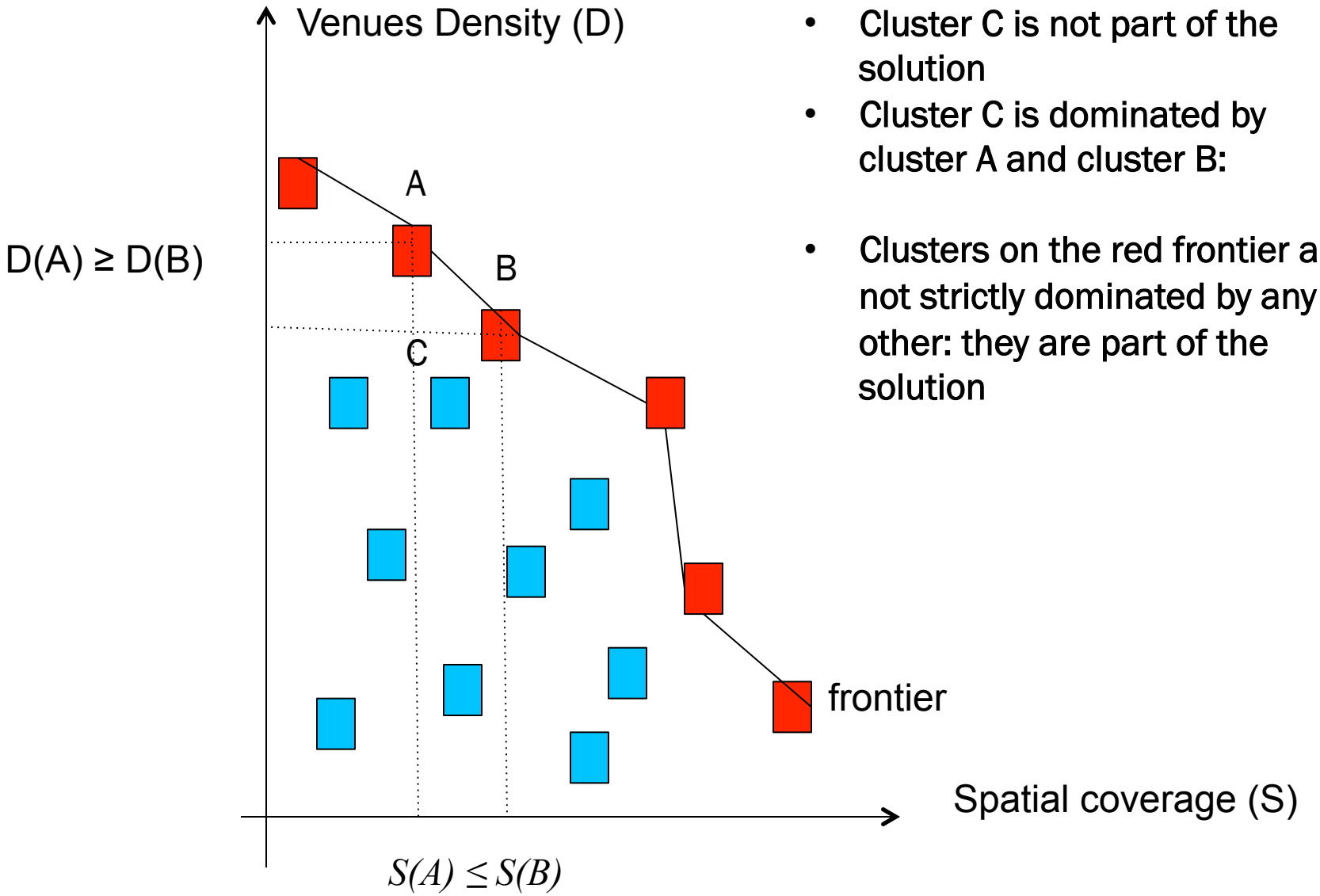
1. **Cluster spatial coverage** (ratio of the BBOX)
   sp_cov = #cells in cluster / #tot_cells;

2. **Venue density (number of venues per unit area)**
   dens = #venues_of_cluster / #cells in cluster;

3. **Category Heterogeneity (fraction of the overall number of categories)**
   het = #categories_of_cluster / #total_categories;

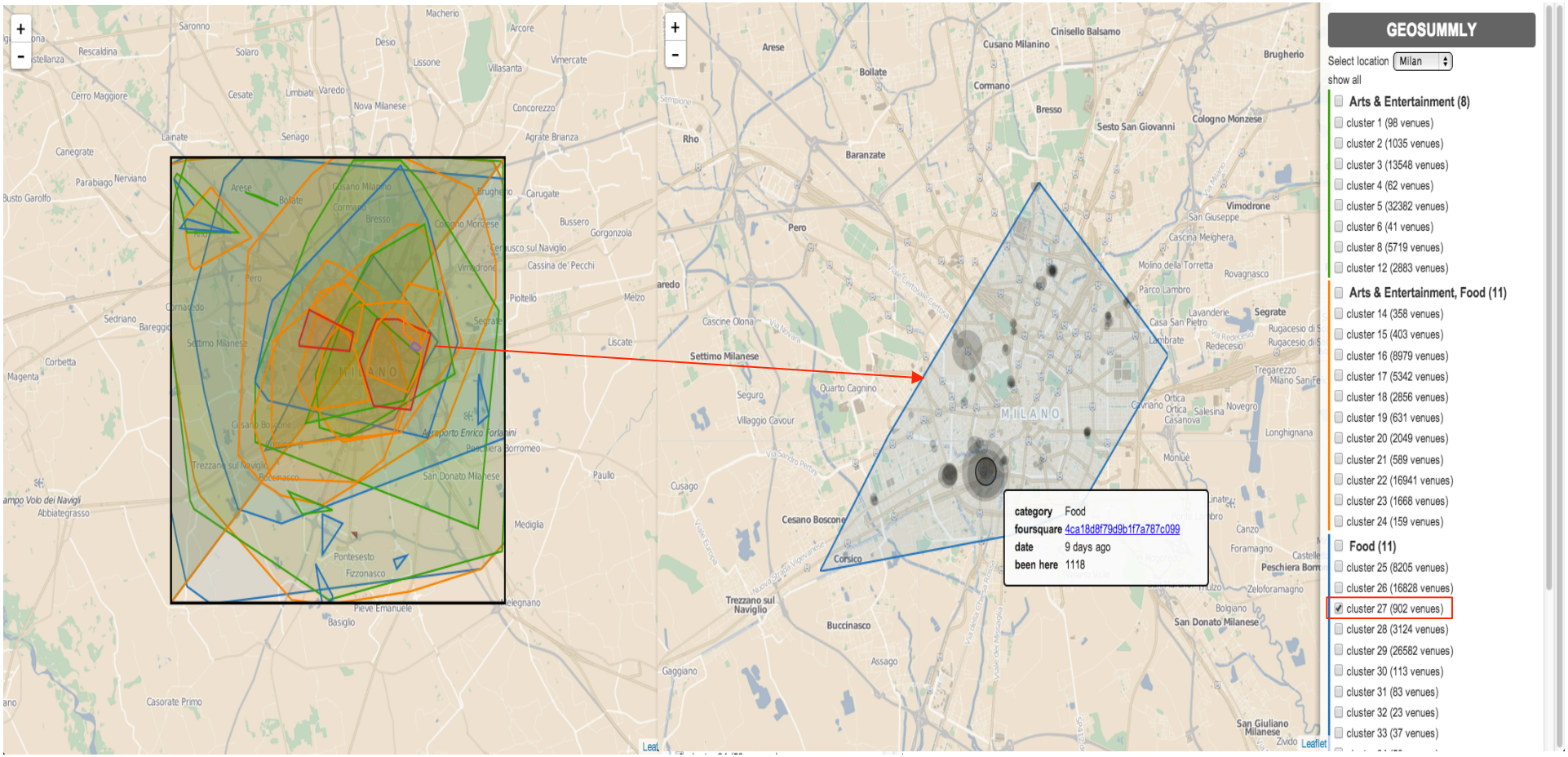
4. **Avg. Distance (between venues in the same cluster)**
   Sum of Squared Errors in Km;

Optimization: Pareto efficiency



- Cluster C is not part of the solution
- Cluster C is dominated by cluster A and cluster B:
- Clusters on the red frontier are not strictly dominated by any other: they are part of the solution

Milan

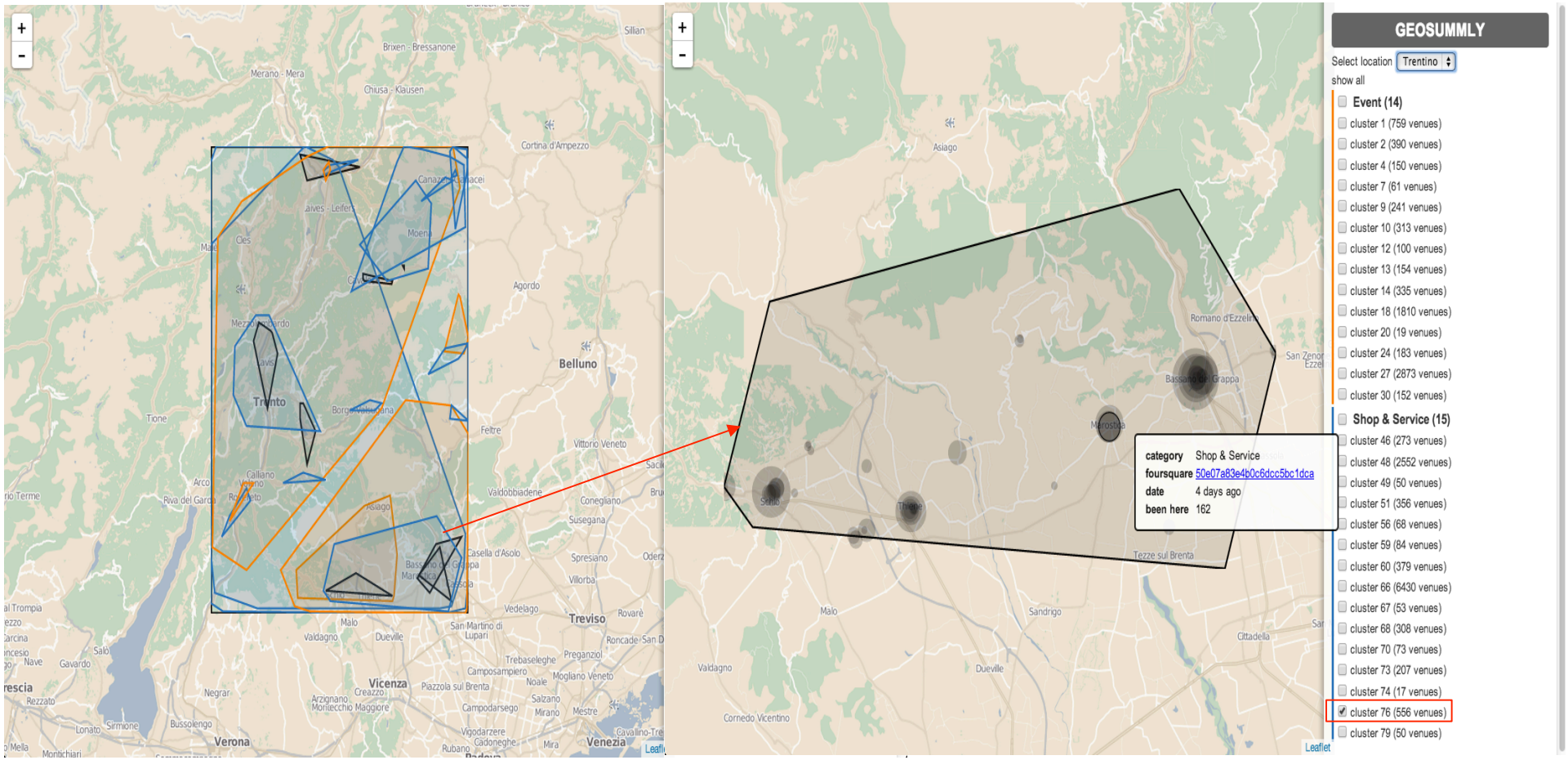


First glance summary

Zooming in “Shop & Service”

SMATF2

Trentino



First glance summary

Zooming in “Shop & Service”

SMATF2

Entity recognition in the semantic web

We applied entity recognition aiming at ontology matching to make data fusion between OpenStreetMap (OSM) and Foursquare (FS)

We performed the following steps:

- Retrieval of a training set by a novel heuristic based on the geographic coordinates and the names similarity of two entities
- The training set contains the pairs of entities from OSM and FS candidate to a match
- Generation of a classifier (based on RandomForest) for each pair of entity classes

Accuracy ranges from 60% to 95% according to the classes

Integration of sources

Foursquare

Event
Shop & Service
Travel & Transport
Professional & Other
Places College & University
Outdoors & Recreation
Residence
Food
Arts & Entertainment
Nightlife Spot

OpenStreetMap

aerialway
barrier
emergency
cycleway
leisure
natural
public_transport
shop
aeroway
building
geological
historic
man_made
office
railway
tourism
amenity
craft
highway
landuse
military
power
route
waterway

SMATF2

SMATF2

Steps for the integration

1. Statistical phase on instances with heuristics

   Geographic coordinates (nearby)

   Similar names of the entities (by cosine similarity)

2. Matching at the level of the category (2nd level)
3. Matching at the level of the instance
   (entities instances classifiers)

Conclusions

- We presented Geosummly,
  - an application based on density based clustering of the venues coming from the social media platforms in a multidimensional space
- Allows the exploitation of social media
- Allows the integration of different platforms
- Creates a new cartography on up-to-date data
- Publishes results in LOD and could exploit it further
- Allows the exploratory analysis on the map for social policy making (e-government), and for the users of the smart city (tourists, citizens, ecc)